Optimised Auto-scaling for Cloud-based Web Service

Jing Jiang

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

UNIVERSITY OF TECHNOLOGY SYDNEY

A thesis submitted for the degree of

Doctor of Philosophy

February 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

Acknowledgements

This thesis is the fruit of my hard work during the whole of my PhD study since being in Australia. It has not been a straight path but it has been an amazing journey in which I have experienced happiness, excitement, depression, struggle and friendship. I feel quite lucky that so many people have offered me a helping hand in this journey. Now is the right time to thank all of them!

First, I would like to express my sincere appreciation and great gratitude to my principal supervisor, Professor Jie Lu, and my co-supervisor, Professor Guangquan Zhang. They are not only mentors in my research life but also the guides in my everyday life, especially in my first year in Australia. In my research work, they offered me ample trust and free space to choose the topic and never doubted my progress, but always made valuable suggestions if I encountered problems.They also guided me on how to express my research results efficiently including how to organise a paper, the right steps to take in writing, and even how to express ideas clearly in sentences. They both always work so

hard and have a strict attitude to research work, which stimulates me to further efforts and from which I will benefit throughout my life. Besides the research work, they were enormously helpful and caring and never said "No" to my requests - for example, they drove to pick up me at the airport on the first day I arrived in Sydney; made suggestions for renting a house; taught me to cook; invited me to their home for celebrations, and even guided me in managing work and life. There are no words to express all my thanks to them; the only way to repay them was to work hard.

I kindly thank all my colleagues in the Decision Systems and e-Service Intelligent (DeSI) lab for their careful participation in my presentation and valuable comments for my research. I also thank many members of the Centre for Quantum Computation and Intelligent Systems (QCIS) and the Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS) for their various assistance and advice which have been of great benefit to this study. I appreciate the financial support from the International Postgraduate Research Scholarship (IPRS) and UTS President's Scholarship (UTSP).

I appreciate the conference and research travel support by Prof Chengqi Zhang, director of QCIS, the UTS Vice Chancellor's conference fund,

and the Faculty of Engineering and Information Technology travel fund.

I sincerely thank Ms. Sue Felix and Ms. Barbara Munday for polishing my English language and doing proofreading for my publications and this thesis; their comments have greatly helped me to improve my academic writing. I would also sincerely thank Ms.Teraesa Ashworth who spent much time introducing me to local culture and helping to improve my spoken English.

I would like to express my earnest appreciation and gratitude to my parents for all their selfless love and generous understanding, also to my brothers and sisters for their warm concern and for looking after our parents when I was not around for all those years.

Last but certainly not least certainly not the least, I feel extremely fortunate that I could share the joy and the pain with my dear husband, Guodong Long. He has been by my side during hard times and has comforted me and encouraged me to continue. I am grateful for his love and endless optimism at times when the barriers seemed impossible to break down.

Abstract

Elasticity and cost-effectiveness are two key features for ensuring that cloud-based web services appeal to more businesses. However, true elasticity and cost-effectiveness in the pay-per-use cloud business model has not yet been fully achieved. The explosion of cloud-based web services brings new challenges to enable the automatic scaling up and down of service provision when the workload is time-varying.

This research studies the problems associated with these challenges. It proposes a novel scheme to achieve optimised auto-scaling for cloud-based web services from three levels of cloud structure: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). At the various levels, auto-scaling for cloud-based web services has different problems and requires different solutions.

At the SaaS level, this study investigates how to design and develop scalable web services, especially for time-consuming applications. To achieve the greatest efficiency, the optimisation of service provision problem is studied by providing the minimum functionality and fastest

scalability performance concerning the speed-up curve and QoS (Quality of Service) of the SLA (Service-Level Agreement). At the PaaS level, this work studies how to support dynamic re-configuration when workloads change and the effective deployment of various kinds of web services to the cloud. To achieve optimised auto-scaling of this deployment, a platform is designed to deploy all web services automatically with the minimal number of cloud resources by satisfying the QoS of SLAs. At the IaaS level for two infrastructure resources of virtual machine (VM) and virtual network (VN), this research focuses on studying two types of cloud-based web service: computation-intensive and bandwidth-intensive. To address the optimised auto-scaling problem for computation-intensive cloud-based web service, data-driven VM auto-scaling approaches are proposed to handle the workload in both stable and dynamic environments. To address the optimised auto-scaling problem for bandwidth-intensive cloud-based web service, this study proposes a novel approach to predict the volume of requests and dynamically adjust the software defined network (SDN)-based network configuration in the cloud to auto-scale the service with minimal cost.

This research proposes comprehensive and profound perspectives to

solve the auto-scaling optimisation problems for cloud-based web services. The proposed approaches not only enable cloud-based web services to minimise resource consumption while auto-scaling service provision to achieve satisfying performance, but also save energy consumption for the global realisation of green computing. The performance of the proposed approaches has been evaluated on a public platform (e.g. Amazon EC2) with the real dataset workload of web services. The experiment results demonstrate that the proposed approaches are practicable and achieve superior performance to other benchmark methods.

Table of Contents

List of Figures

List of Tables